Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date

Larsen, Bjarne; Migicovsky, Zoë; Jeppesen, Anne Aae; Gardner, Kyle M.; Toldam-Andersen, Torben Bo; Myles, Sean; Ørgaard, Marian; Petersen, Mikael Agerlin; Pedersen, Carsten

Published in:
Plant Genome

DOI:
10.3835/plantgenome2018.12.0104

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Many apple cultivars grown today have been frozen in genetic time through clonal propagation for over one hundred years. Consequently, cultivars are not able to adapt to changes in growing conditions or to new pathogens and pests. The industry therefore has to rely on breeding programs to provide new apple cultivars with improved disease resistance and fruit quality under climate changes. Traditional breeding in apple is slow and unpredictable since apple has a long juvenile phase, a high degree of heterozygosity, self-incompatibility, and complex traits controlled by multiple loci (Brown et al., 2003; Kumar et al., 2014). Genomic selection using...
Aldehydes are produced from both pathways and are chain esters and branched chain esters, respectively. Acid pathway or the isoleucine pathway (reviewed by Farneti et al., 2017; Kumar et al., 2015; et al., 2009, 2012; Farneti et al., 2017; Kumar et al., 2015; et al., 2009). Next-generation sequencing (NGS) has resulted in an apple reference genome and thousands of SNPs, and the cost of NGS continues to decrease (Bianco et al., 2016; Bianco et al., 2014; Daccord et al., 2017; Velasco et al., 2010). The challenge, however, is to collect and combine genetic markers with reliable phenotypic data—both quantitative and qualitative. In germplasm collections, historical phenotypic data has been successfully linked with new genomic data and used for genome-wide association studies (GWAS) in apple (Migicovsky et al., 2016), grape (Vitis vinifera L.) (Migicovsky et al., 2017), and barley (Hordeum vulgare L.) (Matthies et al., 2014). However, historical phenotype data from germplasm collections that were not collected for genetic mapping purposes might present a challenge for GWAS, as they may have been collected over a time span of years by different people and different localities (e.g., Toldam-Andersen et al., 2011; Volk and Richards 2008). Today, many collections are being genotyped with high-density marker systems for breeding and management purposes that simultaneously provide data for GWAS. In this study we explore the usefulness of historical phenotypic data from an apple collection for uncovering marker–trait associations through GWAS.

Genetic improvement of fruit quality includes external fruit appearance, such as color, shape, and size, as well as internal traits such as aroma, texture, crispiness, and juiciness (Brown et al., 2003). Despite still being shadowed by major breeding goals, such as disease resistance, fruit quality, and storability, more focus on creating aromatic genotypes with high eating quality should be considered by future breeding programs. However, limited knowledge of the inheritance of important aroma loci impacts and hampers their implementation in MAS as discussed by Dunemann et al. (2009). Apple aroma is a biochemically and genetically complex trait (reviewed by Dixon and Hewett, 2000). More recently, several groups have mapped loci for the content of various aroma volatiles (Dunemann et al., 2009, 2012; Farnetti et al., 2017; Kumar et al., 2015; Ulrich and Dunemann 2012; Vogt et al., 2013).

The largest group of aroma volatiles in apple is esters, which form and contribute to the characteristic fresh, fruity apple flavor. Esters are synthesized via the fatty acid pathway or the isoleucine pathway (reviewed by Schaffer et al., 2007) leading to the formation of straight chain esters and branched chain esters, respectively. Aldehydes are produced from both pathways and are reduced to alcohols via alcohol dehydrogenase. The final step in ester synthesis is catalyzed by alcohol acyl transferases (AATs) (Rowan et al., 1996, 1999).

Other important fruit aroma volatiles include alcohols, aldehydes, ketones, lactones, and terpenoids (El Hadi et al., 2013). Although >350 volatile aroma compounds have been identified in apple (Fuhrmann and Groesch, 2002), a subset of ~20 compounds have been reported to contribute to typical apple aroma and flavor (Dixon and Hewett, 2000). In addition to aroma, sweetness is an important fruit quality trait, which is determined by the three soluble sugars: sucrose, fructose, and glucose. Sucrose is synthesized during photosynthesis and is translocated into sink organs such as fruits. During maturation, enzymatic degradation of sucrose into fructose and glucose is catalyzed by invertases, leading to an increase in these monosaccharides and consequently a decrease in sucrose content (Fuleki et al., 1994; Li et al., 2012; Zhu et al., 2013).

In this work we used a gene bank collection (The Pometum, University of Copenhagen, Denmark) composed mainly of Danish heritage cultivars. We combined data on fruit aroma volatiles, fruit sugars, and 56 historical phenotypic traits with SNPs in a GWAS. Here, we describe significant associations for several fruit esters, sugar content, sugar composition, and harvest date.

MATERIALS AND METHODS

Plant Material and Single Nucleotide Polymorphism Calling

The apple cultivars used throughout this study are mainly Danish heritage cultivars from the gene bank collection The Pometum (University of Copenhagen, Taastrup, Denmark) (Supplemental Table S1). They represent a subset of cultivars used in Larsen et al. (2018). Genotyping-by-sequencing based on Elshire et al. (2011) and subsequent SNP calling and alignment of GBS reads to the ‘Golden Delicious’ genome version 1.0p at Genome Database for Rosaceae (GDR) was performed in a previous study (Larsen et al., 2018) (SNP data available via the Dryad Digital Repository doi:10.5061/dryad.b5s2h4n).

Aroma Volatile Quantification by Dynamic Headspace–Gas Chromatography–Mass Spectrometry

Juice extraction for sugar, acid, and aroma volatile analysis was performed from cultivars collected for The Pometum germplasm collection. Apples were harvested during one growth season. At the trees from which the fruits were harvested, the fruits were thinned manually in early summer after blooming to ensure a uniform distribution of fruits in the canopy and to improve the development of individual fruits. The harvest date decision was based on development in fruit removal force, sensory evaluation (color, softness, smell, and taste), and supported by determination of the Streif Index parameters (percentage soluble solids, firmness, and starch degradation) (Streif, 1996) to process the apples at the most optimal time. Early ripening cultivars, which ripen on the tree, were juiced immediately after harvest. Late ripening cultivars...
that need postharvest ripening for optimal aroma development, were harvested at maturity and initially stored at 4°C for 1 to 2 mo depending on cultivar followed by postharvest ripening at 18°C for 1 to 8 d. The number of days the individual cultivars needed to stay at room temperature to reach an optimal ripe stage varied from cultivar to cultivar and was determined by assessing cultivars individually regarding aroma development and taste. A minimum of 10 well-developed and undamaged apples per cultivar, randomly sampled in the tree canopy, were crushed before juicing with a 20-L hydropress. Juice samples were immediately frozen and kept at −20°C until analysis.

To perform volatile quantification, each 20-mL juice sample was equilibrated to 37°C in a circulating water bath and then purged with nitrogen (100 mL min⁻¹ for 20 min). Volatiles were collected on Tenax-TA traps (200 mg of Tenax-TA, mesh size 60/80, Markes International). After purging, water was removed by applying a flow of dry nitrogen (100 mL min⁻¹ for 10 min). Trapped volatiles were desorbed in an automatic thermal desorption unit (TurboMatrix 350, PerkinElmer). Primary desorption was performed by heating the trap to 250°C with a flow (50 mL min⁻¹) of carrier gas (H2) for 15 min. Stripped volatiles were trapped in a Tenax TA cold trap (30 mg held at 5°C), which was subsequently heated at 300°C for 4 min (secondary desorption, outlet split 1:10). This allowed for rapid transfer of volatiles to a GC-MS (7890A GC-system interfaced with a 5975C VL, MSD with Triple-Axis detector from Agilent Technologies) through a heated (225°C) transfer line. Separation of volatiles was performed on a ZB-Wax capillary column 30 m long by 0.25 mm i.d., 0.50-μm film thicknesses. Column pressure was held constant at 2.3 psi, resulting in an initial flow rate of 1.4 mL min⁻¹ using hydrogen as carrier gas. The column temperature program was as follows: 10 min at 30°C, from 30 to 240°C at 8°C min⁻¹, and finally 5 min at 240°C. The mass spectrometer was operating in the electron ionization mode at 70 eV. Mass-to-charge ratios between 15 and 300 were scanned. Volatile compounds were identified by probability-based matching of their mass spectra with those of a commercial database (Wiley275.L, HP product no. G1035A), and the identifications were confirmed by comparing calculated retention indices with values from online databases (chemspider.com and pherobase.com).

Clustering Aroma Compounds

For clustering analyses, aroma volatile values were log-10 transformed and analyzed with Cluster 3.0 (de Hoon et al., 2004). For hierarchical clustering, distance measures were based on the Pearson correlation (centered) and an average linkage clustering. The data file was visualized as a heat map with dendrograms of both aroma volatiles and cultivars using TreeView (Page, 1996). A principal component analysis of aroma volatile was further performed using the Real Statistics Resource Pack software in Excel (Zaiontz, 2017).

Sugar and Acid Content

Soluble solids and acid content were measured by refractometer (percentage brix) and total acids by titration (with NaOH 0.1 N to pH 8.1), while individual sugars and acids were measured by ion chromatography. Prior to injection, a 1-mL sample was filtered (45-μm syringe filter) and diluted 60× with deionized water. A Metrhom ion-chromatography system (Metrhom AG) equipped with an autosampler (919 IC Plus) ran parallel analyses, with a 20-μL injection to a 881 Compact IC Pro with a 896 Professional detector for the sugars and a 20-μL injection to a 883 Basic IC plus unit for the organic acids. Sugars were separated on a Metrosep CARB 1, 150- by 4.0-mm column at 35°C with a 100 mM NaOH eluent (flow at 1.0 mL min⁻¹). Organic acids were separated on a Metrosep Organic Acid 250- by 7.8-mm column at room temperature with 0.4 mM H₂SO₄ plus 120 mL acetone L⁻¹ as eluent (flow at 0.38 mL min⁻¹). Both columns were installed with a Metrhom guard column.

Historical Phenotypic Traits

Tree and fruit historical phenotypic traits were recorded in a database of The Pometum germplasm collection (Toldam-Andersen et al., 2011). These data were initially available for 64 phenotypic traits. The dataset was pruned to contain only phenotypes with data for >100 cultivars, resulting in a data set of 56 traits (Supplemental Table S2), which were used for GWAS.

Genome-Wide Association Studies

One hundred and forty-five cultivars for which aroma volatile data was available were used for GWAS, which was performed for each recorded aroma volatile using both log-10 transformed VOC data and untransformed data (Supplemental Table S4) and with both general linear model (GLM) and mixed linear model (MLM) and minor allele frequency (MAF) thresholds at 5% resulting in 15802 SNPs or at 10% (11536 SNPs). Sugar and acid quantity data were available for 110 cultivars and historical phenotypic trait data were available for 177 cultivars (Supplemental Table S1). The GWAS were performed using TASSEL 5.0 (Bradbury et al., 2007) both with a GLM and a MLM.

The browser at GDR (https://www.rosaceae.org/) (Jung et al., 2014) was used for aligning and navigating between the two annotated apple genome assemblies: the diploid Golden Delicious genome v. 1.0p at GDR (Velasco et al., 2010), which was used for the SNP-calling; and the haploid Golden Delicious genome (GDDH13) v. 1.1 (Daccord et al., 2017), which was only used for localizing positions of selected SNPs with strong association to the traits. Both genome assemblies were used for searching for candidate genes within a distance of up to 250 kb of SNPs above significance level for Bonferroni corrected threshold at 5.5 (−log-10 to P-value).

Linkage Disequilibrium Decay

Analysis was restricted to 15,802 SNPs with MAF >0.05 and calculated linkage disequilibrium (LD) decay for
all pairwise comparisons within 1000 kb using PLINK (Purcell 2009; Purcell et al., 2007). There are gaps of unknown sequence, represented by 'N', in the apple reference genome v1.0p (Velasco et al., 2010) and to avoid bias in our LD decay measurements, any SNP pairs separated by a large gap (>10,000 'N' designations) were discarded. Results were visualized using the ggplot2 package in R with a locally estimated scatterplot smoothing (LOESS) line (Wickham 2009).

Next, LD was calculated for SNPs within 1000 kb of a SNP with a significant association result for a trait of interest, resulting in 49 to 76 pairwise comparisons. For aroma, the SNP of interest was chr2:1730413; for sugars it was chr1:30129468; for color it was chr9:33085337; and for harvest date it was chr3:31409480. We visualized the relationship between distance from the SNP of interest and LD decay, with the maximum distance ranging from 300 to 1000 kb, by plotting the results using the ggplot2 package in R with a LOESS smoothing line (Wickham 2009).

**DNA Sequencing**

The NAC 18.1 candidate gene (MDP0000868419 in GDR, MD03G1222600 in GDDH13) was amplified by polymerase chain reaction (PCR) from selected apple cultivars using the primers NAC 18.1-F1:CCACCTCCTCAACAGACATTG, NAC 18.1-R1: CCCAATTACAAAGAGCATAACTATATA, and NAC 18.1-F2:AGGCTACGTTTGGAGAGCAA using standard protocols. Products were purified using Exonuclease and Antarctic Phosphatase (New England Biolabs), subsequently sequenced (Eurofins Genomics) and analyzed using CLC Main Workbench (Qiagen Bioinformatics).

**RESULTS**

**Contents of Aroma Volatiles**

We identified 49 fruit VOCs (Supplemental Table S3,S4) performing dynamic headspace GC-MS analysis using apple juice samples from 145 diploid cultivars. Esters were the largest group of volatiles, representing 24 out of 49 observed volatiles, while alcohols and aldehydes were the second largest groups including 12 and seven volatiles, respectively. In addition we identified ketones, terpenes, and phenylpropanoids (Supplemental Table S4). Histograms of the relative VOC amounts (Supplemental Fig. S1) show that most volatiles are having right-skewed distributions. Most juice samples have very low values for most VOCs, though the esters generally have higher values. The ratios between the means and the medians are highly variable and always less than 1 (Supplemental Table S3).

Contents of VOCs on the log-10 scale were clustered as a heatmap (Fig. 1) to visualize the grouping of cultivars according to their aroma profiles. Cooking-type apple cultivars (framed in blue) generally have low amounts
of most aroma volatiles. Namely two groups (framed in red) constituted primarily of dessert apples show particularly high amounts of acetate esters. We performed PCA using VOC contents and found roughly the esters divided into three groups as indicated on Supplemental Fig. S2: \(\text{Acetate esters, methyl and ethyl esters, and butanoate and hexanoate esters.}\) The aldehydes formed two clusters: aldehydes with chain lengths of four to six carbons (butanal, hexanal, and pentanal) clustered, whereas aldehydes with chain lengths of seven to 10 carbons (decanal, heptanal, nonanal, and octanal) cluster together.

**Genome-Wide Association Studies on Aroma Volatiles**

We found a number of marker–trait associations for SNPs above the Bonferroni-corrected significance threshold (Fig. 2, Fig. 3A). These were all studied using quantile–quantile plots (Supplemental Fig. S3) and allele–effect plots as exemplified in Fig. 3B. In some cases, highly right-skewed VOC distributions (Supplemental Fig. S1) produce significant but unreliable associations. Thus, to get better insight into which significant associations that were possibly reliable, we took into consideration the number of associated SNPs forming a peak at a Manhattan plot as well as the shape of the peak and the MAFs of those markers. Manhattan plots based on the MLM and the corresponding quantile–quantile plots for all VOCs are shown in Supplemental Fig. S3. We regard all significant associations in this study as unreliable except for the strong association on chromosome 2, which we investigated further.

On chromosome 2, we found convincing associations for the acetate esters, especially butyl acetate and hexyl acetate (Fig. 2; Fig. 3A). Here, SNPs distributed over a 2- to 3-Mb region showed significant associations; however, most associations were only identified when using a GLM. The scatter plots (Fig. 3B) show the effects of the most significant SNP alleles on the amounts of the two esters. For butyl acetate and hexyl acetate, the cultivars that are heterozygous (CT and AG, respectively) for the loci chr2:1258734 and chr2:1730413 had approximately five times higher ester contents than the homozygous cultivars with genotypes CC and GG, respectively. These differences are highly significant (\(P\)-values \(4.0 \times 10^{-7}\) for butyl acetate and \(8.3 \times 10^{-7}\) for hexyl acetate) (Fig. 3B). We note that only one or two cultivars are homozygous TT and AA. This means that the effect is primarily detected between cultivars with no positive alleles and heterozygous cultivars with one positive allele (Fig. 3B).

In addition to examining VOC content, we examined the ratios between compounds that are linked by a single enzymatic step. This was done to quantify ester formation.
For this purpose we looked at two of the most abundant esters: butyl acetate and hexyl acetate. The precursors of butyl acetate are butanal and butanol, whereas the precursors of hexyl acetate are hexanal and hexanol. We found that all the apple juice samples contain more butyl acetate than butanol and more butanol than butanal. Samples generally had high levels of hexanal and hexyl acetate and lower amounts of hexanol (Supplemental Table S5; Supplemental Fig. S4). Therefore, we calculated the ratios of these compounds for each cultivar (Supplemental Table S5) and performed GWAS on the log-10 transformed and untransformed ratios of butanol/
butanal, butyl acetate/butanol, hexanol/hexanal, and hexyl acetate/hexanol. For both ester/alcohol ratios we found prominent associations in the same region of chromosome 2 as found for acetate esters using log-10 transformed data (Fig. 4; Table 1), while untransformed data gave stronger associations for the butyl acetate/butanol ratio but much less convincing associations for the hexyl acetate/hexanol ratio (Supplemental Fig. S5; Table 1).

To look for candidate genes in this genomic region, the location of the SNPs with the strongest associations was determined in the GDDH13 genome assembly. We found that the alcohol acyl-transferase AAT1 (MD02G1013900) is located within the genomic region, only 163 kb from the SNP with the strongest association (chr2:1730413; Fig. 5).

Sugar and Acid Content

Fructose was the predominant sugar and constituted 55.8% (60.9 g L⁻¹) of the total soluble sugars among all samples. Sucrose constituted in average 32.6% (36.0 g L⁻¹) of the total sugars among all samples, whereas glucose constituted on average 11.4% (12.4 g L⁻¹) of the total soluble sugars (Supplemental Table S6). We found a positive correlation between fructose content and total sugar content ($R^2 = 0.6545$, $P$-value = $1.12 \times 10^{-26}$) as well as between sucrose content and total sugar content ($R^2 = 0.4189$, $P$-value = $2.186 \times 10^{-14}$) but not between glucose content and total sugar content. We also found a weak negative correlation between glucose content and sucrose content (Supplemental Fig. S6).

Table 1. Significant single nucleotide polymorphism (SNP)–trait associations for sugar and harvest time traits with their closest candidate genes (mixed linear model, minor allele frequency [MAF] > 5%).

<table>
<thead>
<tr>
<th>Position in GDR v1.0p</th>
<th>Position in GDDH13</th>
<th>Trait†</th>
<th>Allele</th>
<th>MAF</th>
<th>Variance explained by SNP</th>
<th>$-\log 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr:1:30040671</td>
<td>chr03:28060398</td>
<td>Fructose%</td>
<td>A/G</td>
<td>0.45</td>
<td>33</td>
<td>6.73</td>
</tr>
<tr>
<td>chr:1:30040683</td>
<td>chr03:28060410</td>
<td>Fructose%</td>
<td>C/G</td>
<td>0.45</td>
<td>33</td>
<td>6.73</td>
</tr>
<tr>
<td>chr:1:30040731</td>
<td>chr03:28060461</td>
<td>Fructose%</td>
<td>C/T</td>
<td>0.44</td>
<td>31</td>
<td>6.37</td>
</tr>
<tr>
<td>chr:1:30124669</td>
<td>chr07:32107986</td>
<td>Fructose%</td>
<td>A/G</td>
<td>0.47</td>
<td>41</td>
<td>8.09</td>
</tr>
<tr>
<td>chr:1:30129468</td>
<td>chr07:32121322</td>
<td>Fructose%</td>
<td>C/T</td>
<td>0.45</td>
<td>47</td>
<td>8.94</td>
</tr>
<tr>
<td>chr:1:30151906</td>
<td>chr11:28066346</td>
<td>Fructose%</td>
<td>C/T</td>
<td>0.43</td>
<td>43</td>
<td>8.44</td>
</tr>
<tr>
<td>chr:1:30221387</td>
<td>chr07:32134678</td>
<td>Fructose%</td>
<td>A/C</td>
<td>0.45</td>
<td>43</td>
<td>8.43</td>
</tr>
<tr>
<td>chr:1:31646414</td>
<td>chr01:29191051</td>
<td>Fructose%</td>
<td>A/G</td>
<td>0.49</td>
<td>27</td>
<td>5.64</td>
</tr>
<tr>
<td>chr:1:31643115</td>
<td>chr01:29543175</td>
<td>Fructose%</td>
<td>G/T</td>
<td>0.45</td>
<td>26</td>
<td>5.55</td>
</tr>
<tr>
<td>chr:1:30192468</td>
<td>chr07:32121332</td>
<td>Sucrose content</td>
<td>C/T</td>
<td>0.45</td>
<td>27</td>
<td>5.58</td>
</tr>
<tr>
<td>chr:1:30151906</td>
<td>chr11:28066346</td>
<td>Sucrose content</td>
<td>C/T</td>
<td>0.43</td>
<td>26</td>
<td>5.53</td>
</tr>
<tr>
<td>chr:1:30221387</td>
<td>chr07:32134678</td>
<td>Sucrose%</td>
<td>A/C</td>
<td>0.45</td>
<td>31</td>
<td>6.37</td>
</tr>
<tr>
<td>chr:3:31409362</td>
<td>chr03:30698039</td>
<td>Harvest date</td>
<td>A/C</td>
<td>0.24</td>
<td>28</td>
<td>8.46</td>
</tr>
<tr>
<td>chr:3:31409376</td>
<td>chr03:30698053</td>
<td>Harvest date</td>
<td>C/T</td>
<td>0.24</td>
<td>28</td>
<td>8.46</td>
</tr>
<tr>
<td>chr:3:31409480</td>
<td>chr03:30698157</td>
<td>Harvest date</td>
<td>C/T</td>
<td>0.24</td>
<td>30</td>
<td>8.69</td>
</tr>
<tr>
<td>chr:3:31409362</td>
<td>chr03:30698039</td>
<td>Eating time</td>
<td>A/C</td>
<td>0.24</td>
<td>23</td>
<td>7.51</td>
</tr>
<tr>
<td>chr:3:31409376</td>
<td>chr03:30698053</td>
<td>Eating time</td>
<td>C/T</td>
<td>0.24</td>
<td>23</td>
<td>7.51</td>
</tr>
<tr>
<td>chr:3:31409480</td>
<td>chr03:30698157</td>
<td>Eating time</td>
<td>C/T</td>
<td>0.24</td>
<td>25</td>
<td>7.76</td>
</tr>
</tbody>
</table>

† Fructose% and Sucrose% indicates the content of fructose and sucrose, respectively, in percentage of total sugar content.
The association analyses revealed significant hits for fructose as a percentage of total sugar content, sucrose as a percentage of total sugar content, and for total sucrose content. The significant hits were all in the region spanning ~620 kb on chromosome 1 (Fig. 6; Table 1). This was evident using both the GLM (Supplemental Fig. S7) and the MLM approach (Fig. 6). There was no convincing, significant association for glucose (Fig. 6; Supplemental Fig. S6), and the glucose content is low in general. The allele–effect scatter plots for the most significant SNP showed an additive effect of the T-allele that is positively correlated with low sucrose content and a corresponding high fructose percentage (Fig. 6B). The effect was very significant (Fig. 5B) since cultivars with the TT genotype at chr1:30129468 had an average of 28 g L\(^{-1}\) sucrose in their juice, while cultivars with the CC genotype had nearly twice that value at 50 g L\(^{-1}\) (\(P\)-value \(6.5 \times 10^{-13}\)).

There is a vacuolar invertase, MDP0000149570, described by Hyun et al. (2011) as VIN1, located in this 620-kb region on chromosome 1 in the GDR-v1.0 assembly. The VIN1 is 48 kb from the most significantly associated SNP (chr1:30221387) and we regard this as a good candidate gene for the association. However, VIN1 is not present in the GDR-v1.0p assembly (Supplemental Fig. S8). We identified the corresponding region on chromosome 1 in the most recent version of the apple genome (GDDH13) v. 1.1; the SNPs with the strongest associations were predicted to be located on chromosome 11 and chromosome 7 (Table 1). Using SNPs and gene sequences, we determined that genomic segments of ~35 and 70 kb from chromosome 1 in the initial genome assembly are moved to chromosomes 7 and 11, respectively, in the recent genome assembly (Supplemental Fig. S8). An LD plot of the associated SNPs shows strong LD among all SNPs so they must all be physically linked, and since the flanking SNPs are on chromosome 1 in both genome assemblies, we predict that chromosome 1 is the location of the underlying gene (Supplemental Fig. S9). The SNP closest to VIN1 in the GDDH13 genome is ~15 kb away from the gene.

Acids were composed mainly of malic acid (10.0 g L\(^{-1}\)), while citric acid and succinic acids were present in very small amounts (<0.2 g L\(^{-1}\) on average). We did not detect any significant GWAS associations for any of the acids examined.

Tree and Fruit Phenotypes
The GWAS revealed a very convincing Manhattan peak for fruit color on chromosome 9, which is well known from other studies and validates our GWAS approach for discovering alleles of large effect in apple (Supplemental Fig. S10). We also performed GWAS for harvest date, eating time, and earliness, which are all highly correlated (Supplemental Fig. S11). Where harvest date indicates the ideal harvest time, eating time describes the optimum time to consume the cultivar after postharvest ripening, which is especially relevant for late ripening cultivars. Earliness also describes the ripening time, which is equivalent to eating time, but was measured on a five-step scale. For
all three traits, we found significant associations to the same three SNPs separated by only 14 and 104 bp on chromosome 3 (Fig. 7A; Table 1; Supplemental Fig. S10). These three traits essentially represent three scoring systems of one trait, and thus, the correlation between traits and similar GWAS results are unsurprising.

The harvest date was graded on a 1-to-9 scale reflecting optimal fruit harvest date ranging from late July (1) to late November (9). We found that, on average, there was a 1.5 mo difference in harvest date between cultivars that were homozygous AA vs. CC for SNP chr3:31409362, with heterozygous AC cultivars in-between (Fig. 7B).

The SNP chr3:31409362 is a nonsynonymous change in the gene MDP0000868419 (MD03G1222600 in GDDR13) encoding a NAC transcription factor. We aimed to identify additional putatively causal variants in the coding sequence by sequencing the entire coding sequence in two groups: five AA homozygous cultivars and six CC homozygous cultivars. We found that each group was homozygous for one haplotype; there was no sequence variation within groups. Between groups,
however, there were 18 SNPs and two insertions–deletions (InDels). The two InDels are 3 and 12 bp in length (Supplemental Fig. S12). The SNPs resulted in a total of 12 amino acid substitutions between haplotypes (Supplemental Fig. S13). The 12-bp InDel is in the N-terminal part of the protein and can be explained by duplication of a 6-bp repeat, AGCCGC, in early cultivars (Fig. 8).

**Linkage Disequilibrium Decay**

To determine the extent of LD decay within the germplasm collection, we calculated LD using 34,563 pairwise comparisons for SNPs <160 kb apart with an average inter-SNP distance of ~66 kb (Fig. 8). Among these, there were 3,762 SNP pairs that were <1600 bp apart, with an average distance of ~246 bp (Fig. 9B). The large number of SNPs within a short distance allowed us to determine that genome-wide, LD decays to an $r^2 < 0.2$ within ~200 bp, similar to estimates from American (Migicovsky et al., 2016) and Canadian (McClure et al., 2018) apple collections.

When calculating LD for SNPs located near to a SNP of interest (Fig. 9C–F), we have very few SNPs within 1 kb, making it difficult to assess LD decay with high resolution. For example, the nearest SNP located to the SNP of interest for sugar at chr1:30129468 is 4371 bp away (Supplemental Table S7). In some cases, LD decay within the region does not appear to differ from LD decay on a genome-wide scale. For aroma, the nearest SNP is 12 bp away with $r^2 = 0.87$, while the next SNP is 2060 away with $r^2 = 0.02$. However, our results indicate that, in particular, the LD surrounding the sugar and color SNPs may extend further than expected based on the average genome-wide rate of LD decay (Fig. 9D,E).

**DISCUSSION**

We performed GWAS with 110 to 177 cultivars, which is in the lower limits for what is feasible for association studies. The aroma data are challenging for GWAS because VOC contents vary heavily depending on the maturity stage of the apples. Furthermore, the SNP density is also a limiting factor because of the rapid LD decay in apple. However, we found a strong association for fruit color on chromosome 9 for a known causal gene, indicating that it is possible to identify reliable genotype–phenotype associations using a relatively small number of cultivars. In addition to a small sample size, many of the phenotypic traits examined were historical data, which

![Fig. 7. Manhattan plots using mixed linear model and minor allele frequency >5% for the harvest date. The red, dashed line on (A) indicates the Bonferroni corrected threshold. (B) Allele-effect scatterplot of harvest date in 162 apple cultivars grouped according to their genotype in the single nucleotide polymorphism (SNPs) with the highest associations positioned at 31409362 on chromosome 3. The SNP marker chr3 31409362 explains 23% of the variance in harvest date. Harvest date is recorded on a scale from 1 to 9, where 1 corresponds to late July and 9 corresponds to late November. The differences are statistically significant with P-values at 3.3 × 10^{-8} (AA vs. CA) and 3.6 × 10^{-9} (CA vs. CC). The numbers behind each data point is indicated. The red bars show the median and the gray bars are the first and third quartiles.]

Fig. 8. Alignment of the DNA and protein sequences of the N-terminal of the NAC18.1 transcription factor from apple accessions with the A allele in SNPchr3:31409362 corresponding to a T in position 13 (gene is in reverse orientation) and associated with an early harvest date and accessions with the C allele corresponding to a G in position 13 and associated with a late harvest date. The single nucleotide polymorphisms are highlighted in green and pink and the insertion–deletions in blue and gray.
were not collected for genetic mapping purposes. Thus, we expect strong environmental effects and differences in measurements as a result of the number of observers responsible for phenotyping across years. Despite these limitations, this study identifies several associations of note including aroma, sugar, color, and harvest date.

**Sensory Perception of Juice versus Whole Fruits**

The cultivars Cox’s Orange Pippin and Elstar are important relatives to several cultivars in The Pometum collection studied here (Larsen et al., 2017). The aroma volatiles of the two cultivars have been studied by Fuhrmann and Grosch (2002). The volatile contents were found to be influenced...
by the preparation of the fruit such as whole, unpeeled fruits; fruit flesh; and juice from homogenized apples. When apples were crushed, a rapid enzymatic process started hydrolysis of some esters whereas the content of other esters such as butyl acetate and hexyl acetate rapidly increased after crushing the fruits. As these two esters were actually the two most predominant esters in our study, it is highly likely that the processing of the whole apples into juice has influenced the composition of volatiles. The reason for processing whole apples into juice was to allow maximum release of volatiles for headspace analysis. It should be kept in mind, however, that the volatile content reported here is the volatile content in juice samples, which may differ from the volatile composition in whole apples.

The clustering of VOCs (Fig. 1) reflects that related VOCs, such as acetate esters, are synthesized by the same biosynthetic pathways. It can therefore be expected that such VOCs show associations to the same genomic regions. This was exactly what we observed for acetate esters but also previous work in apples (Rowan et al., 2009) and other fruits (Schwieterman et al., 2014) found similar correlations between VOCs. Keeping in mind that the content of VOCs is only partly controlled by genetics, other factors, such as the growing conditions (Qin et al., 2017) and the exact maturity stage of the fruits (Mehinag et al., 2006), also have a strong influence. The optimal time points for harvest and for juice preparation for each cultivar is thus very important to obtain comparable data. Even smaller differences in the maturity stage at the time of juice preparation and VOC measurements can easily obscure genetic differences. Although much attention in this study was drawn to select the exact optimum harvest time, replications in VOC measurements within and between years would have been highly desirable and could very likely have improved the GWAS results.

**Alcohol Acyl Transferases and Ester Content**

Generally, the juice samples had high amounts of the butyl acetate, lower amounts of butanol, and even lower amounts of butanal (Supplemental Fig. S4). It indicates that butanal is easily converted to butanol and that the alcohol acyl transferases that are responsible for the last step of the ester formation are rate limiting the ester formation (Supplemental Fig. S4). In contrast, we found high levels of hexanal and hexyl acetate but very little hexanol, indicating that formation of hexanol is the rate limiting step. The most prominent association for the acetate/butanol ratio, and to some extent for the hexyl acetate/hexanol ratio, was on chromosome 2 at a position of 0 to 3 Mb.

As alcohol acyl transferases are known to catalyze the final step in the ester biosynthesis, we were searching for known candidate genes involved in ester formation. We found several SNPs distributed over the region from 0 to ~3 Mb in the start of chromosome 2 associated with most acetate esters. Further, these SNPs were only found to be significant using the GLM and not the MLM that corrects for relatedness among samples. This is problematic because rapid LD decay (Migicovsky et al., 2016) suggests that trustful association signals would span over far shorter distances. In addition, the inclusion of a kinship matrix is always desirable in a GWAS to avoid false positive associations resulting from population structure. However, among the apple cultivars used here, 38 parent–offspring relations have been identified forming a significant network from few founder cultivars such as Cox’s Orange Pippin (Larsen et al., 2017, 2018), so the MLM may be overcorrecting and thus masking a biologically meaningful genotype–phenotype association. While the associations for acetate esters on chromosome 2 are promising, the large region of interest makes it challenging to identify exact candidate genes.

The top end of chromosome 2 has been suggested to be a crucial “hot spot” (Kumar et al., 2015) for genes involved in aroma volatile biosynthesis. Here, the gene MdAAT1 (MDP0000214714, MD02G1013900) has been pointed out as the most promising candidate involved in ester biosynthesis. This has been shown both in quantitative trait loci (QTL) studies using segregating populations (Costa et al., 2013; Dunemann et al., 2009, 2012; Rowan et al., 2009; Souleyre et al., 2014; Ulrich and Dunemann 2012; Zini et al., 2005) and in GWAS studies (Amyotte et al., 2017; Farnetti et al., 2017; Kumar et al., 2015). The expression of AAT1 also fits well since it is induced by ethylene and coincides with ester formation in fruits (Schaffer et al., 2007).

We found significant associations for some acetate esters (methyl acetate, pentyl acetate, and propyl acetate), the closest being 163 kb from AAT1. However, in addition to MdAAT1, there are five HXXXX-type acyltransferase encoding genes at the top end of chromosome 2 (Fig. 5), which are all so similar that it is hard to distinguish differences between alleles of the same locus and very similar genes. So, allelic variants obtained by PCR with gene-specific primers, like those described from ‘Royal Gala’ and ‘Granny Smith’ (Souleyre et al., 2014), may in fact represent different paralogous genes of this gene cluster (data not shown). Dunemann et al. (2012) described correlations between AAT1 haplotypes and ester contents but further studies of both gene expression and allelic variants based on RNAseq will be important for clarifying the significance of AAT variants and ultimately allow this information to be used for MAS.

**Vacuolar Invertases and Sugar Types**

The main sugars contributing to sweetness in apple are fructose and sucrose; however, in many studies only the total sugar content is considered. Here we show that by measuring the individual sugars a more detailed GWAS can be performed. Thus, we identified significant associations in the same region on chromosome 1 (Table 1) not only for sucrose content but also for fructose to total sugars ratio and sucrose to total sugars ratio. This is consistent with the findings of Guan et al. (2015) who found the most consistent QTL (ss47583868/ss475876912) for both fructose and sucrose in the same genomic region described here. Sun et al. (2015) also describe QTL for sugar content on chromosome 1 but at another location.
Hyun et al. (2011) described three vacuolar invertase genes including VIN1 (MDP0000149570), which is located on chromosome 1 (position 23870970 to 23875107) in the ‘Golden Delicious’ v1.0 assembly (www.rosaceae.org). This position corresponds approximately to the region from 30160000 to 30170000 on chromosome 1 in the v1.0p genome assembly, which is where our strongest association is. We hypothesize that the gene is located in the target region even though the gene itself could not be detected in v1.0p. The genome assembly GDDH13 v1.1 splits this region of chromosome 1 into three areas on the chromosomes 1, 7, and 11 (Supplemental Fig. S8) and the invertase VIN1 here named MD11G1195800 is located on chromosome 11. However, our work provides evidence that the associated SNP markers all form one strong LD group and therefore are not located on three different chromosomes. Thus, it is likely that VIN1 is located in the peak region and is the best candidate for MAS of sucrose content and sugar ratios but further studies are needed to confirm this.

The invertase gene VIN1 is a strong candidate for the identified association because vacuolar invertases play a key role in sugar metabolism in fruits. It acts by converting sucrose into fructose and glucose in vacuoles where sucrose is pumped in by sucrose transporters and stored as sucrose or as hexoses (Roitsch and Gonzalez, 2004). When more sucrose is split by vacuolar invertase activity, more sucrose may be imported into the vacuole, and it might explain the positive correlation between the content of fructose and total sugars and the fact that cultivars high in fructose are not lower in sucrose (Supplemental Fig. S6). Since fructose is sweeter than sucrose and glucose, an increase in vacuolar invertase activity will result in sweeter apples and the markers identified here will be useful for selecting high-fructose apples.

NAC Transcription Factor and Harvest Date

The three SNPs significantly associated with harvest time in our study were first described by Migicovsky et al. (2016) when examining an American apple collection and by McClure et al. (2018) when examining a Canadian apple collection. Similarly, Urrestarazu et al. (2017) evaluated apples from across Europe and found an association for ripening period in the same genomic area. It is also in accordance with previous reports for QTL for harvest date on chromosome 3 (Kenis et al., 2008; Liebhard et al., 2003). These studies cover a broad range of European and North American germplasm that is distinct from the germplasm studied here, suggesting that this common QTL appears to be general for apple germplasm from Europe and North America.

One of the SNPs (chr3:31409362) is located in the coding region of a NAC transcription factor MDP0000868419 (=MD03G1222600) and causes an amino acid change from tyrosine to aspartic acid. To identify other putatively causal variants, we compared the full-length sequences from six late and five early cultivars and found 12 amino acid substitutions in these two alleles, mostly in positions that are highly conserved among close homologs from other species (Supplemental Fig. S14). The insertion of the peptide glutamine-proline-glutamine-proline in the N-terminal could be of particular interest because it is in a glutamine-rich, variable region that may play an important role for variation among species in this otherwise highly conserved NAC domain. However, this region is not predicted to be directly involved in DNA interactions (Welner et al., 2012) (Supplemental Fig. S15) and further evidence is needed to establish a functional link between these two haplotypes and the speed of ripening in apple. Recently, Moyano et al. (2018) analyzed expression of NAC transcription factors during fruit ripening in strawberry (Fragaria × ananassa Duchesne ex Rozier) using NAC sequences annotated from wild strawberry (F. vesca L.) and found that the closest homolog to the apple NAC18, FaNAC035, is expressed in fruits during ripening. NAC TFs have also been found to be involved in fruit ripening in banana (Musa acuminate Colla), tomato (Solanum lycopersicum L.), and peach [Prunus persica (L.) Batsch] (Ma et al., 2018; Pirona et al., 2013; Shan et al., 2012). Thus, NAC18.1 is a promising functional candidate for fruit ripening in apple that warrants further study.

Linkage Disequilibrium

We used 15,802 SNPs distributed with an average distance of 47 kb according to reported genome size. We find that LD decays within ~200 bp, and thus, it would require at least 100 times more SNPs to saturate the genome for well-powered GWAS. However, for fruit color, we found extended LD decay and SNPs located several hundred kilobases from the causal gene, the R2R3 MYB transcription factor gene (MDP0000259614) previously found to be responsible for fruit color (Ban et al., 2007; Takos et al., 2006), were in significant LD. Extended LD decay was also observed for the sugar ratio. For both of the traits, it may be explained by selection in cultivars for these traits.

CONCLUSION

Notable associations as well as interesting candidate genes were found for harvest date, sugar composition, and content of acetate esters in apple juice. One of the SNPs with strongest association for harvest date is in the coding region of a NAC transcription factor and sequencing alleles from both early and late cultivars showed the existence of two haplotypes with many differences at the amino acid level. The most significant new result from the study is the strong association for type of sugar on chromosome 1 very close to a vacuolar invertase. Future apple breeding might benefit from these genetic markers to perform MAS, decreasing the time and money required for developing and selecting desirable genotypes. Ultimately, once potential candidate genes are confirmed and their function well-characterized, it might be desirable to edit specific alleles to improve specific traits in desired and commercially important cultivars.


